Back to JP200

**Family list**
**1** family member for:
**JP2002334076**
Derived from 1 application.

1 **METHOD FOR PROCESSING TEXT**
Publication info: **JP2002334076 A** - 2002-11-22

Data supplied from the ***esp@cenet*** database - Worldwide

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-334076
(43)Date of publication of application : 22.11.2002

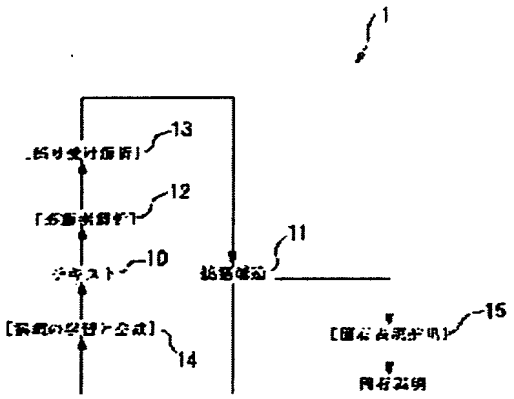
(51)Int.Cl. G06F 17/27
G06F 17/21

(21)Application number : 2001-139563
(22)Date of filing : 10.05.2001

(71)Applicant : COMMUNICATION RESEARCH LABORATORY
(72)Inventor : UCHIMOTO SEIKI
ISAHARA HITOSHI

## (54) METHOD FOR PROCESSING TEXT

(57)Abstract:
PROBLEM TO BE SOLVED: To perform highly accurate text processing by a computer by performing learning on the basis of small learning data in each process included in text processing.
SOLUTION: A text processing method constituted of an analytical process for analyzing syntactic structure and a generation process for generating a text from the syntactic structure is provided with a learning function for repeatedly executing the analytical process and the generation process in constitution including morpheme analysis processing, modification analysis processing and word order learning determination processing and deductively learning regularity at least in any one of the morpheme analysis processing, modification analysis processing and word order learning determination processing.




LEGAL STATUS

| | |
|---|---|
| [Date of request for examination] | 10.05.2001 |
| [Date of sending the examiner's decision of rejection] | 03.12.2002 |
| [Kind of final disposal of application other than the examiner's decision of rejection or application converted registration] | |
| [Date of final disposal for application] | |
| [Patent number] | |
| [Date of registration] | |
| [Number of appeal against examiner's decision of rejection] | 2003-00162 |
| [Date of requesting appeal against examiner's decision of rejection] | 06.01.2003 |
| [Date of extinction of right] | |

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－334076
（P2002－334076A）
(43)公開日　平成14年11月22日(2002.11.22)

| (51)Int.Cl.' | 識別記号 | ＦＩ | | テーマコード'(参考) |
|---|---|---|---|---|
| Ｇ０６Ｆ　17/27 | | Ｇ０６Ｆ　17/27 | Ｊ | ５Ｂ００９ |
| 17/21 | ５５０ | 17/21 | ５５０Ａ | ５Ｂ０９１ |

審査請求　有　　請求項の数７　ＯＬ　（全 14 頁）

(21)出願番号　特願2001－139563(P2001－139563)

(22)出願日　平成13年５月10日(2001.5.10)

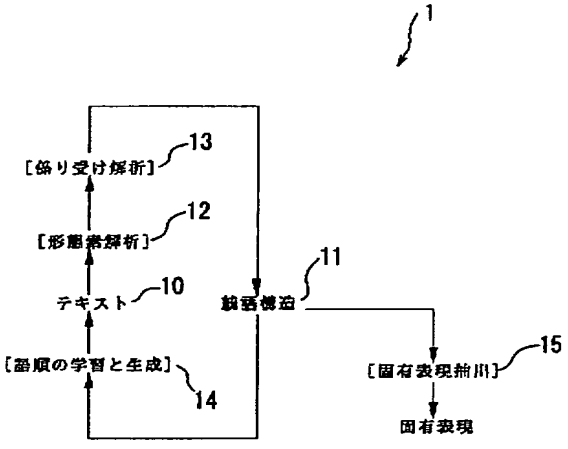
特許法第30条第１項適用申請有り　平成12年11月14日
社団法人人工知能学会主催の「人工知能学会第２種研究会」において文書をもって発表

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　内元　清貴
京都府相楽郡精華町光台２－２－２　独立行政法人通信総合研究所　けいはんな情報通信融合研究センター内
(72)発明者　井佐原　均
京都府相楽郡精華町光台２－２－２　独立行政法人通信総合研究所　けいはんな情報通信融合研究センター内
(74)代理人　100090893
弁理士　渡邊　敏

最終頁に続く

(54)【発明の名称】　テキスト処理方法

(57)【要約】
【課題】　テキスト処理に含まれる各過程で少ない学習データを基に学習を行い、コンピュータによって高精度なテキスト処理を可能にすること。
【解決手段】　統語構造を解析する解析過程と、統語構造からテキストを生成する生成過程とから構成されるテキスト処理方法が、形態素解析処理及び、係り受け解析処理、語順学習決定処理を含む構成において、解析過程と生成過程とを相互に繰り返して実行し、形態素解析処理及び、係り受け解析処理、語順学習決定処理の少なくともいずれかにおける規則性を、演繹的に学習する学習機能を備える。



1

【特許請求の範囲】
【請求項１】言語の解析・生成に関わるコンピュータのテキスト処理方法であって、
該テキスト処理方法が、
統語構造を解析する解析過程と、
統語構造からテキストを生成する生成過程とから構成され、
該解析過程で、
テキストを文法上最小の単位を構成する形態素に分解し、それぞれの形態素に対して文法的属性を決定する形態素解析処理及び、
テキスト内の単数又は連続する複数の形態素からなる文節について、
ある文節が、他のいずれの文節を修飾するかを解析する係り受け解析処理の各処理を含み、
該生成過程で、
言語の語順の学習と決定を行う語順学習決定処理を含む構成において、
解析過程と生成過程とを相互に繰り返して実行し、
形態素解析処理及び、係り受け解析処理、語順学習決定処理の少なくともいずれかにおける規則性を、
演繹的に学習する学習機能を備えたことを特徴とするテキスト処理方法。
【請求項２】前記形態素解析処理が、
テキストから該テキストを構成する文字列の候補を、組み合わせを変えて取り出す構成であって、
取り出した文字列の候補が形態素であるか否か、
又は取り出した文字列の候補の文法的属性が、予め定められた文法的属性群の内のいずれであるかの少なくともいずれかの確率を前記規則性から算出すると共に、
テキストを構成する全ての文字列毎に求められた確率を、互いに積算し、
該積が最大値となる文字列の候補の組み合わせ、又は各形態素の文法的属性の組み合わせの少なくともいずれかを求める方法である請求項１に記載のテキスト処理方法。
【請求項３】前記係り受け解析処理が、
テキストの文末から順に、相対的前方にある前文節と、それより後方にある後文節との２つの文節を、組み合わせを変えて取り出す構成であって、
該前文節が、前文節と該後文節との間にある文節を修飾する関係である確率、
該前文節が、該後文節を修飾する関係である確率、
該前文節が、該後文節よりも後方にある文節を修飾する関係である確率をそれぞれ前記規則性から算出し、
該テキストの各文節に該当する該各確率を、互いに積算することに基づいて係り受け確率を決定する請求項１又は２に記載のテキスト処理方法。
【請求項４】前記係り受け解析処理が、
テキストを構成する全ての文節の組み合わせにおける前記係り受け確率を、
互いに積算し、
該積が最も高くなるように各々の係り受け関係を決定する方法である請求項３に記載のテキスト処理方法。
【請求項５】前記語順学習決定処理において、
テキスト内で、係り受け関係にある文節であって、
該係り文節が２個以上存在する場合に、
該係り文節を２個ずつ抽出して、それらの順序を学習し、
該学習をテキスト内の各文節について行い、
その学習結果を保存する語順モデルを構築する請求項１ないし４に記載のテキスト処理方法。
【請求項６】前記語順学習決定処理において、
テキスト内で、係り受け関係にある文節であって、
該係り文節が２個以上存在する場合に、
該係り文節を２個ずつ抽出して、それらが順序をなす確率を前記語順モデルに基づいて算出すると共に、
全ての係り文節について該確率を求め、
それら全ての確率を互いに積算し、
該積が最大となるような係り文節の順序によって語順を決定する請求項５に記載のテキスト処理方法。
【請求項７】前記解析過程より得られた統語構造から、特定の事物を指す固有表現の抽出を行う請求項１ないし６に記載のテキスト処理方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、日本語等の言語からなるテキストをコンピュータを用いて解析・生成する方法に関するものである。
【０００２】
【従来の技術】コンピュータによって言語のテキストを解析する技術、或いは生成する技術は、言語処理を行う上で必須の技術であり、機械翻訳や、要約システムを実現する上で欠かせない。しかし、言語は曖昧性を有しており、完全な規則性によって構成されるものではないばかりか、自然な言い回しの存在や、語順の自由度の高さなど、コンピュータによって処理を行う際には障害となる問題が非常に多い。そこで、テキスト処理方法については様々な研究がなされている。
【０００３】従来の手法としては、人間によって作成されたテキストを、大量の人手をかけて解析し、該解析に基づいて導かれた規則性をコンピュータに記憶させ、コンピュータは規則性に基づいて、別なテキストを解析・生成する方法がある。しかし、この手法では解析を行うことに膨大な人手とコストを要するばかりでなく、コンピュータは与えられた規則性のみで解析・生成を行うため、人手によって解析された以上の規則性をコンピュータが獲得することがない。そのため、人間が解析した対象テキストに類似のテキストであれば、一定の精度で解析・生成することができるが、別種のテキストの場合に

は、解析精度が低下することがあり、与えられた規則性のみでテキストの解析・生成を行うには限界があった。そして、大量の人手を要さずに容易に実現でき、しかも様々なテキストに対応する高精度なテキスト処理方法は未だ実現されていない。

【０００４】

【発明が解決しようとする課題】本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、テキスト処理に含まれる各過程で少ない学習データを基に学習を行い、コンピュータによって高精度なテキスト処理を可能にすることである。

【０００５】

【課題を解決するための手段】本発明は、上記の課題を解決するために、次のような情報埋込方法を創出する。すなわち、言語の解析・生成に関わるコンピュータのテキスト処理方法であって、該テキスト処理方法が、統語構造を解析する解析過程と、統語構造からテキストを生成する生成過程とから構成される。該解析過程では、テキストを文法上最小の単位を構成する形態素に分解し、それぞれの形態素に対して文法的属性を決定する形態素解析処理及び、テキスト内の単数又は連続する複数の形態素からなる文節について、ある文節が、他のいずれの文節を修飾するかを解析する係り受け解析処理の各処理を含む。また、該生成過程では、言語の語順の学習と決定を行う語順学習決定処理を含む。本構成において、解析過程と生成過程とを相互に繰り返して実行し、形態素解析処理及び、係り受け解析処理、語順学習決定処理の少なくともいずれかにおける規則性を、演繹的に学習する学習機能を備える。

【０００６】前記形態素解析処理が、テキストから該テキストを構成する文字列の候補を、組み合わせを変えて取り出す構成であって、取り出した文字列の候補が形態素であるか否か、又は取り出した文字列の候補の文法的属性が、予め定められた文法的属性群の内のいずれであるかの少なくともいずれかの確率を前記規則性から算出する。そして、テキストを構成する全ての文字列毎に求められた確率を、互いに積算し、該積が最大値となる文字列の候補の組み合わせ、又は各形態素の文法的属性の組み合わせの少なくともいずれかを求め、形態素解析処理を行ってもよい。

【０００７】前記係り受け解析処理が、テキストの文末から順に、相対的前方にある前文節と、それより後方にある後文節との２つの文節を、組み合わせを変えて取り出す構成であって、該前文節が、前文節と該後文節との間にある文節を修飾する関係である確率、該前文節が、該後文節を修飾する関係である確率、該前文節が、該後文節よりも後方にある文節を修飾する関係である確率をそれぞれ前記規則性から算出し、該テキストの各文節に該当する該各確率を、互いに積算することに基づいて係り受け確率を決定してもよい。そして、前記係り受け解析処理が、テキストを構成する全ての文節の組み合わせにおける前記係り受け確率を、互いに積算し、該積が最も高くなるように各々の係り受け関係を決定する方法であってもよい。

【０００８】前記語順学習決定処理において、テキスト内で、係り受け関係にある文節であって、該係り文節が２個以上存在する場合に、該係り文節を２個ずつ抽出して、それらの順序を学習し、該学習をテキスト内の各文節について行い、その学習結果を保存する語順モデルを構築してもよい。さらに、上記の場合に、係り文節を２個ずつ抽出して、それらが順序をなす確率を前記語順モデルに基づいて算出すると共に、全ての係り文節について該確率を求め、それら全ての確率を互いに積算し、該積が最大となるような係り文節の順序によって語順を決定するテキスト処理方法でもよい。

【０００９】前記解析過程より得られた統語構造から、特定の事物を指す固有表現の抽出を行ってもよい。

【００１０】

【発明の実施の形態】以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。以下においては、テキストの１例として、日本語によるテキストを挙げて説述するが、本発明の実施方法は、性質上実現出来ない場合を除き、いかなる言語に対しても適用可能である。図１に本発明におけるテキスト処理方法（１）の説明図を示す。

【００１１】ここで、テキスト処理とはテキスト（１０）を解析し、そこから統語構造（１１）を得る、あるいは、統語構造（１１）からテキスト（１０）を生成する処理のことである。本発明においては、統語構造（１１）を解析する解析過程と、統語構造（１１）からテキスト（１０）を生成する生成過程とを循環的に行うことを特徴とし、解析過程には形態素解析（１２）及び、係り受け解析（１３）の各処理を含み、生成過程には語順の学習生成処理（１４）を含む。さらに、統語構造（１１）から意味解析過程である固有表現抽出（１５）処理を行い、該処理において固有表現の学習・抽出を可能としている。

【００１２】このようにテキストと統語構造とを関連付ける処理が可能となることにより、様々な応用が期待される。例えば、これらの処理により得られた統語構造を日本語以外の対象言語の統語構造へマッピングすることにより、翻訳が可能となるし、得られた統語構造から重要な部分だけを残して生成することにより、テキストの要約が可能となる。また、意味解析によって得られた固有表現は、情報抽出のための重要な基礎情報であるだけでなく、形態素解析、構文解析にフィードバックすることにより、より高精度の解析結果を得るための手掛かりとなり得る情報である。以下、各処理について詳述する。

【００１３】初めに、本発明における各処理で採用する最大エントロピーモデル（以下、ＭＥモデルと呼ぶ。）につき説述する。ＭＥモデルでは、文脈、すなわち観測される情報は、素性と呼ばれる個々の要素によって表される。そして、１個の文がある素性を満たすか否かを表す２値関数を導入する。該２値関数を用い、素性が既知のテキスト中に現れる期待値が、未知なテキスト中においても変わらないという制約のもと、文が生起する確率を推定する。そして、各々の素性には、学習に用いるデータにおける確率分布のエントロピーが最大になるように重み付けを行う。このエントロピーを最大にするという操作によって、既知データに観測されなかったような素性、或いは稀にしか観測されなかった素性については、それぞれの出力値に対して確率値が等確率になるように、或いは近付くように、重み付けされる。以上によって、ＭＥモデルによる確率分布は、素性を引数とする関数として表される。

【００１４】一般に確率モデルでは、文脈、すなわち観測される情報と、そのときに得られる出力値との関係は既知のデータから推定される確率分布によって表される。いろいろな状況に対してできるだけ正確に出力値を予測するためには文脈を細かく定義する必要があるが、細かくしすぎると既知のデータにおいてそれぞれの文脈に対応する事例の数が少なくなりデータが疎らになる問題、すなわちデータスパースネスの問題が生じる。

【００１５】しかし、ＭＥモデルにおいては、上記のように未知のデータに対して考慮した重み付けがなされるため上記データスパースネスの問題に効果的に対応することができる。すなわち、ＭＥモデルは例えば言語現象などのように既知データにすべての現象が現れ得ないような現象を扱うのに適したモデルであり、本発明では、該モデルをテキスト処理における各処理過程に採用している。

【００１６】本発明におけるテキストから統語構造を導出する解析過程に、ＭＥモデルを適用する実施例を次に示す。まず、形態素解析処理についてその方法を説述する。図２に、「先生になった」というテキストを形態素解析する事例を示す。ここで形態素解析の形態素とは、単語や接辞など、文法上、最小の単位となる要素のことである。そして、形態素解析とは、与えられた文を形態素の並びに分解し、それぞれの形態素に対し文法的属性、例えば品詞や活用などを決定する処理のことである。例えば、上記の例によると、「先生」、「に」、「なった」がそれぞれ形態素として見出し語に分類され、それぞれに読みや基本形と共に、文法的属性が付与される。

【００１７】従来の形態素解析において問題となっているのは、辞書に登録されていない、あるいは学習に用いるテキストに現れないが形態素となり得る単語（以下、未知語と呼ぶ。）をどのように扱うかということである。この未知語の問題に対処するため、従来は大きく２つの方法がとられている。その１つは未知語を自動獲得し、辞書に登録する方法であり、もう１つは未知語でも解析できるようなモデルを作成する方法である。本実施例では、この両者の利点を生かすため、前者の方法で獲得した単語を辞書に登録し、後者のモデルにその辞書を利用できる仕組みを取り入れている。そして、これらの手法をＭＥモデルによって実現することにより、辞書の情報を学習する機構を容易に組み込めるだけでなく、字種や字種変化などの情報を用いて学習に用いるテキストから未知語の性質を学習することもできるようになった。

【００１８】本実施例ではＭＥモデルに適用するために、形態素としての尤もらしさを確率として表す。すなわち、文が与えられたとき、その文を形態素解析するという問題は文を構成する各文字列に、２つの識別符号のうち１つ、つまり、形態素であるか否かを示す「１」又は「０」を割り当てる問題に置き換えることができる。さらに、形態素である場合には文法的属性を付与するために「１」を文法的属性の数だけ分割する。すると、文法的属性の数がｎ個のとき、各文字列に「０」から「ｎ」までのうちいずれかの識別符号を割り当てる問題に置き換えることができる。

【００１９】したがって、本実施例における形態素解析にＭＥモデルを用いた手法では、文字列が、形態素であって、かついずれかの文法的属性を持つとしたときの尤もらしさを前記ＭＥモデルにおける確率分布の関数に適用することで求められる。形態素解析においてはこの尤もらしさを表す確率に、規則性を見い出すことで処理を行っている。用いる素性としては、着目している文字列の字種の情報、その文字列が辞書に登録されているかどうか、１つ前の形態素からの字種の変化、１つ前の形態素の品詞などの情報を用いる。１個の文が与えられたとき、文全体で確率の積が最大になるよう形態素に分割し文法的属性を付与する。最適解の探索には適宜公知のアルゴリズムを用いることができる。なお、用いる素性は任意に変更可能である。

【００２０】本発明における形態素解析にＭＥモデルを用いた手法は、従来からの未知語の問題に効果的に対応することができる。たとえば、形態素等を詳細に解析済みのあるテキストを用いた実験では、全形態素に対して区切りと品詞を正しく推定できた割合が約９６％という高精度な結果を得ている。また、実験により、辞書の精度に及ぼす影響の大きさ、および、本手法が、固有名詞、人名、組織名、地名など未知語になりやすいものに対して比較的に推定精度がよいことが分かっている。

【００２１】さらに解析過程においては、係り受け解析にも、ＭＥモデルによる解析手法を取り入れている。次にこの点につき詳述する。どの文節がどの文節を修飾するかという日本語の係り受け関係には、主に以下の特徴

があるとされている。すなわち、
（１）係り受けは前方から後方に向いている。
（２）係り受け関係は交差しない。（以下、これを非交差条件と呼ぶ。）
（３）係り要素は受け要素を１つだけもつ。
（４）ほとんどの場合、係り先の決定には前方の文脈を必要としない。
本実施例では、これらの特徴に着目し、統計的手法と文末から文頭に向けて解析する方法を組み合わせることにより高い解析精度を得ることを実現した。

【００２２】本手法では、文末から順に２つずつ文節を取り上げ、それらが係り受けの関係にあるかどうかを統計的に決定する。その際、文節あるいは文節間にみられる情報を素性として利用するが、どのような素性を利用するかが精度に影響する。文節は、前の主辞にあたる部分と後ろの助詞や活用形にあたる部分に分けて考え、それぞれの素性とともに文節間の距離や句読点の有無なども素性として考慮した。さらに括弧の有無や文節間の助詞「は」の有無、係り側の文節と同じ助詞や活用形が文節間にもあるか否か、素性間の組み合わせについても考慮している。

【００２３】ＭＥモデルによればこういった様々な素性を扱うことができる。そして、この方法では決定木や最尤推定法などを用いた従来の手法に比べて学習データの大きさが１０分の１程度であるにも関わらず、同程度以上の精度が得られる。この手法は学習に基づくシステムとして、最高水準の精度を得られる手法である。さらに、本実施例ではさらに高精度化を図るため、次の手法を取り入れている。すなわち、従来は、学習データから得られる情報を基に、２つの文節が係り受け関係にあるか否かを予測するのに有効な素性を学習していたが、本実施例では、新たに前文節が「後文節を越えて先にある文節に係る」「後文節に係る」「後文節との間にある文節に係る」の３つの状態のどれであるかを予測するのに有効な情報を学習するシステムを開発した。

【００２４】次に、実際にこのモデルから係り受け確率がどのように求まるかを示す。図３に、ある文節（一番左の文節）より後方に５つの文節がある場合に、係り先の候補となる各文節との関係における確率を示す。図中で、「越える」（３１）は上記「後文節を越えて先にある文節に係る」を表し、「係る」（３２）は「後文節に係る」、「間」（３３）は「後文節との間にある文節に係る」に対応する。なお、本発明で言う規則性はこれら確率に表れる。図４は、各候補に係る係り受け確率を求める実施例である。このシステムでは文末から文頭に向かって解析するため、ある文節より後方の文節については、破線の矢印で表されるような係り受け関係がすでに決まったものとして説述する。候補１に係る係り受け確率の算出を例に採ると、候補１が係り先であり、候補１は候補２に、さらに候補５に係る。一方候補３は別個に候補４に係り、さらに候補５に係る。

【００２５】この場合の係り元の文節に関する係り受け確率は、次のように求める。すなわち、候補３及び４は独立した係り受け関係であって、その確率は１とすることができ、候補１に係る確率は図３より０．４であって、候補１は係り元と、候補２及び候補５との間にあるので、各確率は、それぞれ０．１、０．６となる。これをそれぞれ積算し、平方根をとることで、係り受け確率を算出する。同様に、各候補について算出するが、このとき、候補３と候補４は上記非交差条件を満たさないために、この文節の係り先の候補とはなり得ない。ＭＥモデルを用いた係り受け解析では、１個の文全体の確率はそれぞれの文節について求めた係り受け確率の積で表され、非交差条件を満足する条件下で、その積の値が最も高くなるように各々の係り受けを決めることになる。

【００２６】以上、統語構造を解析する解析過程における形態素解析と、係り受け解析にＭＥモデルを用いた実施形態を示した。本発明においては、これらを必ずしも用いる場合に限らず、任意の解析手法を用いることができる。また、形態素解析や係り受け解析を含む限り、さらに他の解析処理を含んでも構わない。

【００２７】次に、生成過程における語順の学習生成過程につき、ＭＥモデルを用いた手法を示す。日本語は語順が自由であると言われている。しかし、これまでの言語学的な調査によると実際には、時間を表す副詞の方が主語より前に来やすい、長い修飾句を持つ文節は前に来やすいといった何らかの傾向がある。もしこの傾向をうまく整理することができれば、それは自然な文を生成する際に有効な情報となる。ここで語順とは、係り相互間の語順、つまり同じ文節に係っていく文節の順序関係を意味するものとする。語順を決定する要因にはさまざまなものがあり、例えば、修飾句の長い文節は短い文節より前に来やすい、「それ」などの文脈指示語を含む文節は前に来やすい、などがあげられる。

【００２８】本発明においては、上記のような要素と語順の傾向との関係、すなわち規則性を所定のテキストから学習する手法を考案した。この手法では、語順の決定にはどの要素がどの程度寄与するかだけでなく、どのような要素の組み合わせのときにどのような傾向の語順になるかということも学習に用いるテキストから演繹的に学習することができる。個々の要素の寄与の度合はＭＥモデルを用いて効率良く学習する。係り文節の数によらず２つずつ取り上げてその順序を学習する。

【００２９】１つの実施例として、学習に用いるテキストに「昨日／太郎は／テニスを／した。」（／は文節の区切りを表す。）という文があった場合を考える。動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の３つである。このうち２文節ずつ、つまり「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の３つのペアを取り上げ、それぞれ

この語順が適切であると仮定して学習する。素性としては文節の持つ属性などを考える。例えば、「昨日／太郎は／した。」という関係からは「時相名詞」の方が「固有名詞」より前に来るという情報、「太郎は／テニスを／した。」という関係からは「は」格の方が「を」格より前に来るという情報などを用いる。

【００３０】文を生成する際には、この学習したモデルを用いて、係り受け関係にある文節を入力とし、その係り文節の順序を決めることができる。語順の決定は次の手順で行なう。まず、係り文節について可能性のある並びをすべて考える。次に、それぞれの並びについて、その係り文節の順序が適切である確率を学習したモデルを用いて求める。この確率は、順序が適切であるか否かの「０」または「１」に置き換え、前記ＭＥモデルにおける確率分布の関数に適用することで求められる。そして、全体の確率が最大となる並びを解とする。全体の確率は、係り文節を２つずつ取り上げたときその順序が適切である確率を計算し、それらの積として求める。例えば、前記「昨日／太郎は／テニスを／した。」という文において、動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の３つである。この３つの係り文節の順序を以下の手順で決定する。

【００３１】図５に係り文節の順序が適切である確率の計算例を示す。まず、２個の文節ずつ、すなわち「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の３つの組み合わせを取り上げ、学習した規則性によりそれぞれこの語順が適切である各確率を求める。例えば、図において「昨日」「太郎は」の語順になる確率は「ｐ*（昨日，太郎は）」で表され、その確率は０．６とする。同様に、「昨日」「テニスを」は０．８、「太郎は」「テニスを」は０．７とすると、図５における１段目の語順（５１）の確率は各確率を積算し、０．３３６となる。次に、６つの語順（５１ないし５６）の可能性すべてについて全体の確率を計算し、最も確率の高いもの「昨日／太郎は／テニスを／した。」（５１）が最も適切な語順であるとする。

【００３２】学習されたモデルの性能は、そのモデルを用いて語順を決めるテストを行ない、元の文における語順とどの程度一致するかを調べることによって定量的に評価することができる。学習したモデル、すなわち規則性を用いて語順を決定させたとき、元のテキストと一致する割合は、前記の解析済みテキストを使用した実験で約７５％であった。さらに、一致しなかった語順においても、その半数はモデルを用いて決定した語順でも不自然ではなく、本発明において効果的な語順の学習・生成が可能であることが示されている。

【００３３】最後に、本発明においては、上記一連の解析過程及び生成過程に加え、意味解析システムを備える。すなわち、意味解析システムの１つとして、本発明において、固有名詞で表されるような特定の事物を指す固有表現を学習により自動抽出する固有表現抽出処理（１５）のシステムを作成する。固有表現として抽出するのは、「特許庁」のように組織の名称を表すもの、「川端康成」のように人名を表すもの、「神戸」のように地名を表すもの、「スペースシャトル」のように固有物の名称を表すものおよび、「９月２８日」、「午後３時」、「１００万円」、「１０％」のように日付、時間、金銭、割合を表す表現である。

【００３４】抽出方法は、以下の通りである。

（１）テキストを単語（正確には形態素）に分割して品詞を割り当てる。例えば、「兵庫県内」は「兵庫（名詞）／県内（名詞）」のように分割される。

（２）各固有表現ごとに固有表現の始まり、中間、終り、単独を表す識別符号（以下、ラベルと呼ぶ。）を用意しておき、演繹的に学習した規則性に基づいて各々の単語に対し付与するべきラベルを推定する。ラベルの推定にはＭＥモデルを用いている。例えば、「兵庫（名詞）／県内（名詞）」は「兵庫＜地名：単独＞／県内＜ラベルなし＞」のように推定される。推定に用いる情報は、着目している単語を含み前後２単語ずつ合計５単語に関する見出し語、品詞の情報である。各ラベルの尤もらしさを確率として計算し、１個の文全体における確率の積の値が高くなり、かつラベルとラベルの間の連接規則を満たすように付与するラベルを決める。１個の文における最適解の探索には各処理段階における最適解をすべて保持する公知のアルゴリズムを用いていることができる。

（３）システムがよく生じる誤りについてその誤りを訂正する書き換え規則を予め規則性の１つとして用意しておき、これを後処理に用いる。例えば、「兵庫＜地名：単独＞／県内＜ラベルなし＞」は「兵庫県＜地名：単独＞／内＜ラベルなし＞」のように書き換えられる。

（４）最後にこの結果から「兵庫県」を地名として抽出する。

本発明における手法によると、人間のパフォーマンスの９割程度の精度で固有表現を抽出でき、従来に比して効果的な固有表現の抽出が可能となった。

【００３５】以上のように本発明では、解析から生成に亙るテキスト処理を、最大エントロピーモデルを用いた学習という一貫した枠組みで処理をしている。そして、解析過程、すなわち形態素解析（単語の切り出し、品詞推定）、係り受け解析や、固有表現抽出を行う意味解析システムから、生成（語順の学習と決定）に至るまでの各処理を、予め解析済みのテキストを用いた学習によって実現する。さらにそれらを繰り返して実行することによって、少ない学習データにもかかわらず、大量の人手をかけて作成される規則に基づく方法に近い精度を実現でき、コストの抑制だけでなく、幅広い文章に対応可能なテキスト処理方法を提供することができる。これら技術は、自動翻訳技術や、テキストの要約技術に用いるだ

けでなく、例えば、コンピュータにおけるかな漢字変換等、いかなる言語処理にも適用することが可能である。

【００３６】

【発明の効果】本発明は、以上の構成を備えるので、次の効果を奏する。請求項１に記載のテキスト処理方法によると、解析過程及び生成過程を互いに繰り返して実行することによって、学習を行う解析済みテキストが少ない場合であっても、効果的に学習を行うことができ、高精度なテキスト処理方法を提供することができる。これによって、コストの低廉化と共に、高機能化を図ることができる。

【００３７】請求項２に記載のテキスト処理方法によると、形態素解析にＭＥモデルを適用することができるので、請求項１に記載の循環的な学習に好適であり、コンピュータにおける処理に馴染みやすい。これによって、本発明におけるテキスト処理方法はより高精度化を図ることができ、処理の高速化にも寄与する。

【００３８】請求項３に記載のテキスト処理方法によると、係り受け確率を定数的に求めることができるので、より高精度な係り受け関係を導出することができ、ひいては高精度なテキスト処理方法に奉仕する。

【００３９】請求項４に記載のテキスト処理方法によると、１個の文全体について全ての係り受け関係の確率を求めるので、文全体として最適な係り受け関係を導出することができ、高精度な係り受け解析が可能となる。これにより高精度なテキスト処理方法に寄与する。

【００４０】請求項５に記載のテキスト処理方法によると、学習によって語順モデルを構築するので、学習を行う解析済みテキストが少ない場合であっても、効果的に学習を行うことができ、高精度なテキスト処理方法を提＊供することができる。

【００４１】請求項６に記載のテキスト処理方法によると、請求項５の方法により構築された語順モデルを用いることができるので、最適な語順の決定を効果的に行うことができる。

【００４２】請求項７に記載のテキスト処理方法によると、固有表現の抽出処理を行うので、形態素解析の精度向上に寄与し、ひいては高精度なテキスト処理方法が実現できる。

【図面の簡単な説明】

【図１】本発明によるテキスト処理方法の説明図

【図２】形態素解析の説明図

【図３】係り受け確率の算出実施例における各確率一覧図

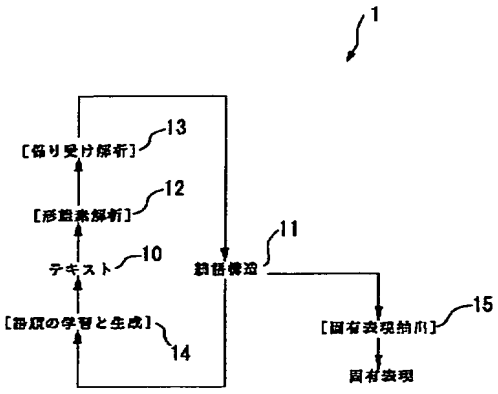
【図４】係り受け確率の算出実施例

【図５】語順の学習生成における順序が適切である確率の計算例

【符号の説明】

１　テキスト処理方法
１０　テキスト
１１　統語構造
１２　形態素解析処理
１３　係り受け解析処理
１４　語順の学習生成処理
１５　固有表現抽出処理
３１　後文節を越えて先にある文節に係る確率
３２　後文節に係る確率
３３　後文節との間にある文節に係る確率
５１ないし５６　係り文節の語順の並べ替え例

【図１】



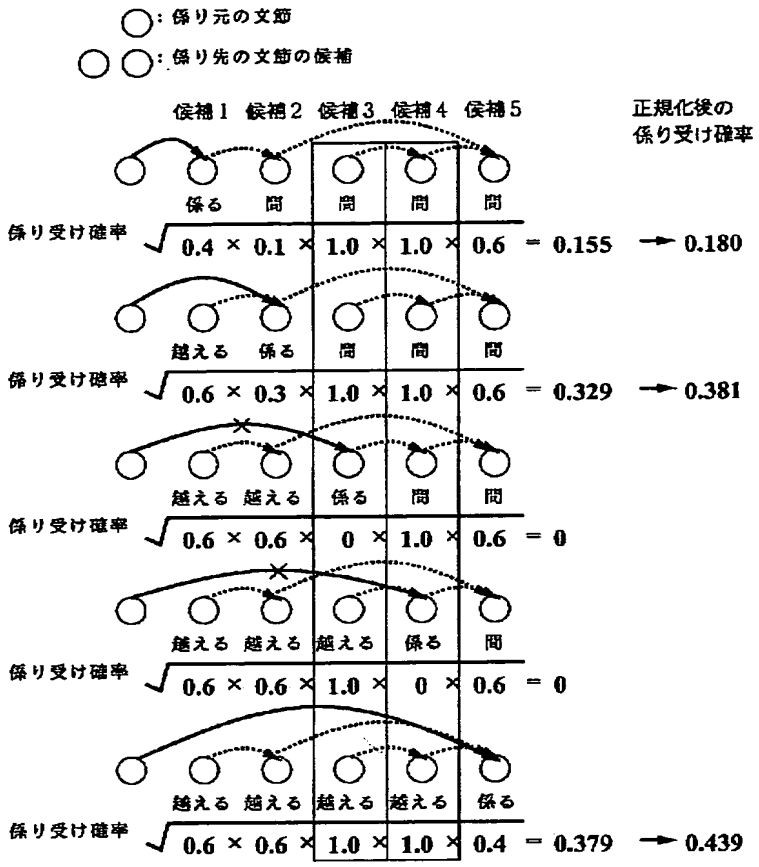

【図２】

| 見出し語 | 読み | 基本形 | 品詞 | 活用型 | 活用形 |
|---|---|---|---|---|---|
| 先生 | (せんせい) | 先生 | 普通名詞 | | |
| に | (に) | に | 格助詞 | | |
| なった | (なった) | なる | 動詞 | 子音動詞ラ行 | タ形 |

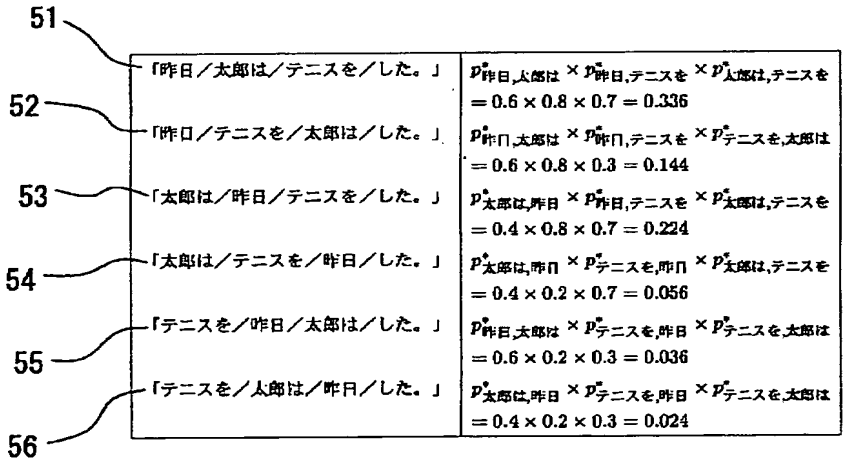
【図３】

| | 31 | 32 | 33 |
|---|---|---|---|
| 候補 | 越える | 係る | 間 |
| 候補１ | 0.6 | 0.4 | 0 |
| 候補２ | 0.6 | 0.3 | 0.1 |
| 候補３ | 0.3 | 0.5 | 0.2 |
| 候補４ | 0.1 | 0.5 | 0.4 |
| 候補５ | 0 | 0.4 | 0.6 |

【図４】


○：係り元の文節
○ ○：係り先の文節の候補
候補１ 候補２ 候補３ 候補４ 候補５
正規化後の係り受け確率
係る 間 間 間 間
係り受け確率 0.4 × 0.1 × 1.0 × 1.0 × 0.6 = 0.155 → 0.180
越える 係る 間 間 間
係り受け確率 0.6 × 0.3 × 1.0 × 1.0 × 0.6 = 0.329 → 0.381
越える 越える 係る 間 間
係り受け確率 0.6 × 0.6 × 0 × 1.0 × 0.6 = 0
越える 越える 越える 係る 間
係り受け確率 0.6 × 0.6 × 1.0 × 0 × 0.6 = 0
越える 越える 越える 越える 係る
係り受け確率 0.6 × 0.6 × 1.0 × 1.0 × 0.4 = 0.379 → 0.439

【図５】


51
「昨日／太郎は／テニスを／した。」
$p^*_{昨日,太郎は} \times p^*_{昨日,テニスを} \times p^*_{太郎は,テニスを}$
$= 0.6 \times 0.8 \times 0.7 = 0.336$
52
「昨日／テニスを／太郎は／した。」
$p^*_{昨日,太郎は} \times p^*_{昨日,テニスを} \times p^*_{テニスを,太郎は}$
$= 0.6 \times 0.8 \times 0.3 = 0.144$
53
「太郎は／昨日／テニスを／した。」
$p^*_{太郎は,昨日} \times p^*_{昨日,テニスを} \times p^*_{太郎は,テニスを}$
$= 0.4 \times 0.8 \times 0.7 = 0.224$
54
「太郎は／テニスを／昨日／した。」
$p^*_{太郎は,昨日} \times p^*_{テニスを,昨日} \times p^*_{太郎は,テニスを}$
$= 0.4 \times 0.2 \times 0.7 = 0.056$
55
「テニスを／昨日／太郎は／した。」
$p^*_{昨日,太郎は} \times p^*_{テニスを,昨日} \times p^*_{テニスを,太郎は}$
$= 0.6 \times 0.2 \times 0.3 = 0.036$
56
「テニスを／太郎は／昨日／した。」
$p^*_{太郎は,昨日} \times p^*_{テニスを,昨日} \times p^*_{テニスを,太郎は}$
$= 0.4 \times 0.2 \times 0.3 = 0.024$

【手続補正書】
【提出日】平成14年7月26日（2002．7．26）
【手続補正1】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正内容】
【書類名】　明細書
【発明の名称】　テキスト処理方法
【特許請求の範囲】
【請求項1】言語の解析・生成に関わるコンピュータのテキスト処理方法であって、
該テキスト処理方法が、
統語構造を解析する解析過程と、
統語構造からテキストを生成する生成過程とから構成され、
該解析過程で、
テキストを文法上最小の単位を構成する形態素に分解し、それぞれの形態素に対して文法的属性を決定する形態素解析処理及び、
テキスト内の単数又は連続する複数の形態素からなる文節について、
ある文節が、他のいずれの文節を修飾するかを解析する係り受け解析処理の各処理を含み、
該生成過程で、
言語の語順の学習と決定を行う語順学習決定処理を含む構成において、
解析過程と生成過程とを相互に繰り返して実行することにより、
形態素解析処理においては、テキストから該テキストを構成する文字列の候補を、組み合わせを変えて取り出し、該取り出した文字列の候補が、形態素であって、かついずれかの文法的属性を持つとしたときの尤もらしさを表す形態素尤度確率、又は、
係り受け解析処理においては、該ある文節が、係り先の候補となる各文節との関係における確率、
語順学習決定処理においては、係り受け関係にあるテキスト内の全ての係り文節の並びについて、その係り文節の順序が適切である確率、
の少なくともいずれかを、
最大エントロピーモデルを用いて学習する学習機能を備えたことを特徴とするテキスト処理方法。
【請求項2】前記形態素解析処理が、
前記形態素尤度確率を前記最大エントロピーモデルにより算出すると共に、
テキストを構成する全ての文字列毎に求められた確率を、互いに積算し、
該積が最大値となる文字列の候補の組み合わせ、又は各形態素の文法的属性の組み合わせの少なくともいずれかを求める方法である請求項1に記載のテキスト処理方法。
【請求項3】前記係り受け解析処理が、
テキストの文末から順に、相対的前方にある前文節と、それより後方にある後文節との2つの文節を、組み合わせを変えて取り出す構成であって、
該前文節が、前文節と該後文節との間にある文節を修飾する関係である確率、該前文節が、該後文節を修飾する関係である確率、
該前文節が、該後文節よりも後方にある文節を修飾する関係である確率をそれぞれ前記最大エントロピーモデルにより算出し、
該テキストの各文節に該当する該各確率を、互いに積算することに基づいて係り受け確率を決定する請求項1又は2に記載のテキスト処理方法。
【請求項4】前記係り受け解析処理が、
テキストを構成する全ての文節の組み合わせにおける前記係り受け確率を、
互いに積算し、
該積が最も高くなるように各々の係り受け関係を決定する方法である請求項3に記載のテキスト処理方法。
【請求項5】前記語順学習決定処理において、
テキスト内で、係り受け関係にある文節であって、
該係り文節が2個以上存在する場合に、
該係り文節を2個ずつ抽出して、それらの順序を前記最大エントロピーモデルを用いて学習し、
該学習をテキスト内の各文節について行い、
その学習結果を保存する語順モデルを構築する請求項1ないし4に記載のテキスト処理方法。
【請求項6】前記語順学習決定処理において、
テキスト内で、係り受け関係にある文節であって、
該係り文節が2個以上存在する場合に、
該係り文節を2個ずつ抽出して、それらが順序をなす確率を前記語順モデルに基づいて算出すると共に、
全ての係り文節について該確率を求め、
それら全ての確率を互いに積算し、
該積が最大となるような係り文節の順序によって語順を決定する請求項5に記載のテキスト処理方法。
【請求項7】前記解析過程より得られた統語構造から、特定の事物を指す固有表現の抽出を行う請求項1ないし6に記載のテキスト処理方法。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、日本語等の言語からなるテキストをコンピュータを用いて解析・生成する方法に関するものである。
【0002】
【従来の技術】コンピュータによって言語のテキストを解析する技術、或いは生成する技術は、言語処理を行う

上で必須の技術であり、機械翻訳や、要約システムを実現する上で欠かせない。しかし、言語は曖昧性を有しており、完全な規則性によって構成されるものではないばかりか、自然な言い回しの存在や、語順の自由度の高さなど、コンピュータによって処理を行う際には障害となる問題が非常に多い。そこで、テキスト処理方法については様々な研究がなされている。

【０００３】従来の手法としては、人間によって作成されたテキストを、大量の人手をかけて解析し、該解析に基づいて導かれた規則性をコンピュータに記憶させ、コンピュータは規則性に基づいて、別なテキストを解析・生成する方法がある。しかし、この手法では解析を行うことに膨大な人手とコストを要するばかりでなく、コンピュータは与えられた規則性のみで解析・生成を行うため、人手によって解析された以上の規則性をコンピュータが獲得することがない。そのため、人間が解析した対象テキストに類似のテキストであれば、一定の精度で解析・生成することができるが、別種のテキストの場合には、解析精度が低下することがあり、与えられた規則性のみでテキストの解析・生成を行うには限界があった。そして、大量の人手を要さずに容易に実現でき、しかも様々なテキストに対応する高精度なテキスト処理方法は未だ実現されていない。

【０００４】

【発明が解決しようとする課題】本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、テキスト処理に含まれる各過程で少ない学習データを基に学習を行い、コンピュータによって高精度なテキスト処理を可能にすることである。

【０００５】

【課題を解決するための手段】本発明は、上記の課題を解決するために、次のようなテキスト生成方法を創出する。すなわち、言語の解析・生成に関わるコンピュータのテキスト処理方法であって、該テキスト処理方法が、統語構造を解析する解析過程と、統語構造からテキストを生成する生成過程とから構成される。該解析過程では、テキストを文法上最小の単位を構成する形態素に分解し、それぞれの形態素に対して文法的属性を決定する形態素解析処理及び、テキスト内の単数又は連続する複数の形態素からなる文節について、ある文節が、他のいずれの文節を修飾するかを解析する係り受け解析処理の各処理を含む。また、該生成過程では、言語の語順の学習と決定を行う語順学習決定処理を含む。本構成において、解析過程と生成過程とを相互に繰り返して実行し、最大エントロピーモデルを用いて学習する学習機能を備える。最大エントロピーモデルを用いて学習するのは、形態素解析処理においては、テキストから該テキストを構成する文字列の候補を、組み合わせを変えて取り出し、該取り出した文字列の候補が、形態素であって、かついずれかの文法的属性を持つとしたときの尤もらしさを表す形態素尤度確率、又は、係り受け解析処理においては、該ある文節が、係り先の候補となる各文節との関係における確率、語順学習決定処理においては、係り受け関係にあるテキスト内の全ての係り文節の並びについて、その係り文節の順序が適切である確率の少なくともいずれかである。

【０００６】前記形態素解析処理が、前記形態素尤度確率を前記最大エントロピーモデルにより算出すると共に、テキストを構成する全ての文字列毎に求められた確率を、互いに積算し、該積が最大値となる文字列の候補の組み合わせ、又は各形態素の文法的属性の組み合わせの少なくともいずれかを求め、形態素解析処理を行ってもよい。

【０００７】前記係り受け解析処理が、テキストの文末から順に、相対的前方にある前文節と、それより後方にある後文節との２つの文節を、組み合わせを変えて取り出す構成であって、該前文節が、前文節と該後文節との間にある文節を修飾する関係である確率、該前文節が、該後文節を修飾する関係である確率、該前文節が、該後文節よりも後方にある文節を修飾する関係である確率をそれぞれ前記最大エントロピーモデルにより算出し、該テキストの各文節に該当する該各確率を、互いに積算することに基づいて係り受け確率を決定してもよい。そして、前記係り受け解析処理が、テキストを構成する全ての文節の組み合わせにおける前記係り受け確率を、互いに積算し、該積が最も高くなるように各々の係り受け関係を決定する方法であってもよい。

【０００８】前記語順学習決定処理において、テキスト内で、係り受け関係にある文節であって、該係り文節が２個以上存在する場合に、該係り文節を２個ずつ抽出して、それらの順序を前記最大エントロピーモデルを用いて学習し、該学習をテキスト内の各文節について行い、その学習結果を保存する語順モデルを構築してもよい。さらに、上記の場合に、係り文節を２個ずつ抽出して、それらが順序をなす確率を前記語順モデルに基づいて算出すると共に、全ての係り文節について該確率を求め、それら全ての確率を互いに積算し、該積が最大となるような係り文節の順序によって語順を決定するテキスト処理方法でもよい。

【０００９】前記解析過程より得られた統語構造から、特定の事物を指す固有表現の抽出を行ってもよい。

【００１０】

【発明の実施の形態】以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。以下においては、テキストの１例として、日本語によるテキストを挙げて説述するが、本発明の実施方法は、性質上実現出来ない場合を除き、いかなる言語に対しても適用可能である。図１に本発明におけるテキスト処理方法（１）の説明図を示す。

【００１１】ここで、テキスト処理とはテキスト（１０）を解析し、そこから統語構造（１１）を得る、あるいは、統語構造（１１）からテキスト（１０）を生成する処理のことである。本発明においては、統語構造（１１）を解析する解析過程と、統語構造（１１）からテキスト（１０）を生成する生成過程とを循環的に行うことを特徴とし、解析過程には形態素解析（１２）及び、係り受け解析（１３）の各処理を含み、生成過程には語順の学習生成処理（１４）を含む。さらに、統語構造（１１）から意味解析過程である固有表現抽出（１５）処理を行い、該処理において固有表現の学習・抽出を可能としている。

【００１２】このようにテキストと統語構造とを関連付ける処理が可能となることにより、様々な応用が期待される。例えば、これらの処理により得られた統語構造を日本語以外の対象言語の統語構造へマッピングすることにより、翻訳が可能となるし、得られた統語構造から重要な部分だけを残して生成することにより、テキストの要約が可能となる。また、意味解析によって得られた固有表現は、情報抽出のための重要な基礎情報であるだけでなく、形態素解析、構文解析にフィードバックすることにより、より高精度の解析結果を得るための手掛かりとなり得る情報である。以下、各処理について詳述する。

【００１３】初めに、本発明における各処理で採用する最大エントロピーモデル（以下、ＭＥモデルと呼ぶ。）につき説述する。ＭＥモデルでは、文脈、すなわち観測される情報は、素性と呼ばれる個々の要素によって表される。そして、１個の文がある素性を満たすか否かを表す２値関数を導入する。該２値関数を用い、素性が既知のテキスト中に現れる期待値が、未知なテキスト中においても変わらないという制約のもと、文が生起する確率を推定する。そして、各々の素性には、学習に用いるデータにおける確率分布のエントロピーが最大になるように重み付けを行う。このエントロピーを最大にするという操作によって、既知データに観測されなかったような素性、或いは稀にしか観測されなかった素性については、それぞれの出力値に対して確率値が等確率になるように、或いは近付くように、重み付けされる。以上によって、ＭＥモデルによる確率分布は、素性を引数とする関数として表される。

【００１４】一般に確率モデルでは、文脈、すなわち観測される情報と、そのときに得られる出力値との関係は既知のデータから推定される確率分布によって表される。いろいろな状況に対してできるだけ正確に出力値を予測するためには文脈を細かく定義する必要があるが、細かくしすぎると既知のデータにおいてそれぞれの文脈に対応する事例の数が少なくなりデータが疎らになる問題、すなわちデータスパースネスの問題が生じる。

【００１５】しかし、ＭＥモデルにおいては、上記のように未知のデータに対して考慮した重み付けがなされるため上記データスパースネスの問題に効果的に対応することができる。すなわち、ＭＥモデルは例えば言語現象などのように既知データにすべての現象が現れ得ないような現象を扱うのに適したモデルであり、本発明では、該モデルをテキスト処理における各処理過程に採用している。

【００１６】本発明におけるテキストから統語構造を導出する解析過程に、ＭＥモデルを適用する実施例を次に示す。まず、形態素解析処理についてその方法を説述する。図２に、「先生になった」というテキストを形態素解析する事例を示す。ここで形態素解析の形態素とは、単語や接辞など、文法上、最小の単位となる要素のことである。そして、形態素解析とは、与えられた文を形態素の並びに分解し、それぞれの形態素に対し文法的属性、例えば品詞や活用などを決定する処理のことである。例えば、上記の例によると、「先生」、「に」、「なった」がそれぞれ形態素として見出し語に分類され、それぞれに読みや基本形と共に、文法的属性が付与される。

【００１７】従来の形態素解析において問題となっているのは、辞書に登録されていない、あるいは学習に用いるテキストに現れないが形態素となり得る単語（以下、未知語と呼ぶ。）をどのように扱うかということである。この未知語の問題に対処するため、従来は大きく２つの方法がとられている。その１つは未知語を自動獲得し、辞書に登録する方法であり、もう１つは未知語でも解析できるようなモデルを作成する方法である。本実施例では、この両者の利点を生かすため、前者の方法で獲得した単語を辞書に登録し、後者のモデルにその辞書を利用できる仕組みを取り入れている。そして、これらの手法をＭＥモデルによって実現することにより、辞書の情報を学習する機構を容易に組み込めるだけでなく、字種や字種変化などの情報を用いて学習に用いるテキストから未知語の性質を学習することもできるようになった。

【００１８】本実施例ではＭＥモデルに適用するために、形態素としての尤もらしさを確率として表す。すなわち、文が与えられたとき、その文を形態素解析するという問題は文を構成する各文字列に、２つの識別符号のうち１つ、つまり、形態素であるか否かを示す「１」又は「０」を割り当てる問題に置き換えることができる。さらに、形態素である場合には文法的属性を付与するために「１」を文法的属性の数だけ分割する。すると、文法的属性の数がｎ個のとき、各文字列に「０」から「ｎ」までのうちいずれかの識別符号を割り当てる問題に置き換えることができる。

【００１９】したがって、本実施例における形態素解析にＭＥモデルを用いた手法では、文字列が、形態素であって、かついずれかの文法的属性を持つとしたときの尤

もらしさを前記ＭＥモデルにおける確率分布の関数に適用することで求められる。形態素解析においてはこの尤もらしさを表す確率に、規則性を見い出すことで処理を行っている。用いる素性としては、着目している文字列の字種の情報、その文字列が辞書に登録されているかどうか、１つ前の形態素からの字種の変化、１つ前の形態素の品詞などの情報を用いる。１個の文が与えられたとき、文全体で確率の積が最大になるよう形態素に分割し文法的属性を付与する。最適解の探索には適宜公知のアルゴリズムを用いることができる。なお、用いる素性は任意に変更可能である。

【００２０】本発明における形態素解析にＭＥモデルを用いた手法は、従来からの未知語の問題に効果的に対応することができる。たとえば、形態素等を詳細に解析済みのあるテキストを用いた実験では、全形態素に対して区切りと品詞を正しく推定できた割合が約９６％という高精度な結果を得ている。また、実験により、辞書の精度に及ぼす影響の大きさ、および、本手法が、固有名詞、人名、組織名、地名など未知語になりやすいものに対して比較的に推定精度がよいことが分かっている。

【００２１】さらに解析過程においては、係り受け解析にも、ＭＥモデルによる解析手法を取り入れている。次にこの点につき詳述する。どの文節がどの文節を修飾するかという日本語の係り受け関係には、主に以下の特徴があるとされている。すなわち、

（１）係り受けは前方から後方に向いている。

（２）係り受け関係は交差しない。（以下、これを非交差条件と呼ぶ。）

（３）係り要素は受け要素を1 つだけもつ。

（４）ほとんどの場合、係り先の決定には前方の文脈を必要としない。

本実施例では、これらの特徴に着目し、統計的手法と文末から文頭に向けて解析する方法を組み合わせることにより高い解析精度を得ることを実現した。

【００２２】本手法では、文末から順に２つずつ文節を取り上げ、それらが係り受けの関係にあるかどうかを統計的に決定する。その際、文節あるいは文節間にみられる情報を素性として利用するが、どのような素性を利用するかが精度に影響する。文節は、前の主辞にあたる部分と後ろの助詞や活用形にあたる部分に分けて考え、それぞれの素性とともに文節間の距離や句読点の有無なども素性として考慮した。さらに括弧の有無や文節間の助詞「は」の有無、係り側の文節と同じ助詞や活用形が文節間にもあるか否か、素性間の組み合わせについても考慮している。

【００２３】ＭＥモデルによればこういった様々な素性を扱うことができる。そして、この方法では決定木や最尤推定法などを用いた従来の手法に比べて学習データの大きさが１０分の１程度であるにも関わらず、同程度以上の精度が得られる。この手法は学習に基づくシステムとして、最高水準の精度を得られる手法である。さらに、本実施例ではさらに高精度化を図るため、次の手法を取り入れている。すなわち、従来は、学習データから得られる情報を基に、２つの文節が係り受け関係にあるか否かを予測するのに有効な素性を学習していたが、本実施例では、新たに前文節が「後文節を越えて先にある文節に係る」「後文節に係る」「後文節との間にある文節に係る」の３つの状態のどれであるかを予測するのに有効な情報を学習するシステムを開発した。

【００２４】次に、実際にこのモデルから係り受け確率がどのように求まるか示す。図３に、ある文節（一番左の文節）より後方に５つの文節がある場合に、係り先の候補となる各文節との関係における確率を示す。図中で、「越える」（３１）は上記「後文節を越えて先にある文節に係る」を表し、「係る」（３２）は「後文節に係る」、「間」（３３）は「後文節との間にある文節に係る」に対応する。　図４は、各候補に係る係り受け確率を求める実施例である。このシステムでは文末から文頭に向かって解析するため、ある文節より後方の文節については、破線の矢印で表されるような係り受け関係がすでに決まったものとして説述する。候補１に係る係り受け確率の算出を例に採ると、候補１が係り先であり、候補１は候補２に、さらに候補５に係る。一方候補３は別個に候補４に係り、さらに候補５に係る。

【００２５】この場合の係り元の文節に関する係り受け確率は、次のように求める。すなわち、候補３及び４は独立した係り受け関係であって、その確率は１とすることができ、候補１に係る確率は図３より０．４であって、候補１は係り元と、候補２及び候補５との間にあるので、各確率は、それぞれ０．１、０．６となる。これをそれぞれ積算し、平方根をとることで、係り受け確率を算出する。同様に、各候補について算出するが、このとき、候補３と候補４は上記非交差条件を満たさないために、この文節の係り先の候補とはなり得ない。ＭＥモデルを用いた係り受け解析では、１個の文全体の確率はそれぞれの文節について求めた係り受け確率の積で表され、非交差条件を満足する条件下で、その積の値が最も高くなるように各々の係り受けを決めることになる。

【００２６】以上、統語構造を解析する解析過程における形態素解析と、係り受け解析にＭＥモデルを用いた実施形態を示した。本発明においては、これらを必ずしも用いる場合に限らず、任意の解析手法を用いることができる。また、形態素解析や係り受け解析を含む限り、さらに他の解析処理を含んでも構わない。

【００２７】次に、生成過程における語順の学習生成過程につき、ＭＥモデルを用いた手法を示す。日本語は語順が自由であると言われている。しかし、これまでの言語学的な調査によると実際には、時間を表す副詞の方が主語より前に来やすい、長い修飾句を持つ文節は前に来やすいといった何らかの傾向がある。もしこの傾向をう

まく整理することができれば、それは自然な文を生成する際に有効な情報となる。ここで語順とは、係り相互間の語順、つまり同じ文節に係っていく文節の順序関係を意味するものとする。語順を決定する要因にはさまざまなものがあり、例えば、修飾句の長い文節は短い文節より前に来やすい、「それ」などの文脈指示語を含む文節は前に来やすい、などがあげられる。

【００２８】本発明においては、上記のような要素と語順の傾向との関係、すなわち規則性を所定のテキストから学習する手法を考案した。この手法では、語順の決定にはどの要素がどの程度寄与するかだけでなく、どのような要素の組み合わせのときにどのような傾向の語順になるかということも学習に用いるテキストから学習することができる。個々の要素の寄与の度合はＭＥモデルを用いて効率良く学習する。係り文節の数によらず２つずつ取り上げてその順序を学習する。

【００２９】１つの実施例として、学習に用いるテキストに「昨日／太郎は／テニスを／した。」（／は文節の区切りを表す。）という文があった場合を考える。動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の３つである。このうち２文節ずつ、つまり「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の３つのペアを取り上げ、それぞれこの語順が適切であると仮定して学習する。素性としては文節の持つ属性などを考える。例えば、「昨日／太郎は／した。」という関係からは「時相名詞」の方が「固有名詞」より前に来るという情報、「太郎は／テニスを／した。」という関係からは「は」格の方が「を」格より前に来るという情報などを用いる。

【００３０】文を生成する際には、この学習したモデルを用いて、係り受け関係にある文節を入力とし、その係り文節の順序を決めることができる。語順の決定は次の手順で行なう。まず、係り文節について可能性のある並びをすべて考える。次に、それぞれの並びについて、その係り文節の順序が適切である確率を学習したモデルを用いて求める。この確率は、順序が適切であるか否かの「０」または「１」に置き換え、前記ＭＥモデルにおける確率分布の関数に適用することで求められる。そして、全体の確率が最大となる並びを解とする。全体の確率は、係り文節を２つずつ取り上げたときその順序が適切である確率を計算し、それらの積として求める。例えば、前記「昨日／太郎は／テニスを／した。」という文において、動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の３つである。この３つの係り文節の順序を以下の手順で決定する。

【００３１】図５に係り文節の順序が適切である確率の計算例を示す。まず、２個の文節ずつ、すなわち「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の３つの組み合わせを取り上げ、ＭＥモデルによりそれぞれこの語順が適切である各確率を求める。例えば、図において「昨日」「太郎は」の語順になる確率は「ｐ＊（昨日，太郎は）」で表され、その確率は０．６とする。同様に、「昨日」「テニスを」は０．８、「太郎は」「テニスを」は０．７とすると、図５における１段目の語順（５１）の確率は各確率を積算し、０．３３６となる。次に、６つの語順（５１ないし５６）の可能性すべてについて全体の確率を計算し、最も確率の高いもの「昨日／太郎は／テニスを／した。」（５１）が最も適切な語順であるとする。

【００３２】学習されたモデルの性能は、そのモデルを用いて語順を決めるテストを行ない、元の文における語順とどの程度一致するかを調べることによって定量的に評価することができる。学習したモデル、すなわち規則性を用いて語順を決定させたとき、元のテキストと一致する割合は、前記の解析済みテキストを使用した実験で約７５％であった。さらに、一致しなかった語順においても、その半数はモデルを用いて決定した語順でも不自然ではなく、本発明において効果的な語順の学習・生成が可能であることが示されている。

【００３３】最後に、本発明においては、上記一連の解析過程及び生成過程に加え、意味解析システムを備える。すなわち、意味解析システムの１つとして、本発明において、固有名詞で表されるような特定の事物を指す固有表現を学習により自動抽出する固有表現抽出処理（１５）のシステムを作成する。固有表現として抽出するのは、「特許庁」のように組織の名称を表すもの、「川端康成」のように人名を表すもの、「神戸」のように地名を表すもの、「スペースシャトル」のように固有物の名称を表すものおよび、「９月２８日」、「午後３時」、「１００万円」、「１０％」のように日付、時間、金銭、割合を表す表現である。

【００３４】抽出方法は、以下の通りである。

（１）テキストを単語（正確には形態素）に分割して品詞を割り当てる。例えば、「兵庫県内」は「兵庫（名詞）／県内（名詞）」のように分割される。

（２）各固有表現ごとに固有表現の始まり、中間、終り、単独を表す識別符号（以下、ラベルと呼ぶ。）を用意しておき、学習結果に基づいて各々の単語に対し付与するべきラベルを推定する。ラベルの推定にはＭＥモデルを用いている。例えば、「兵庫（名詞）／県内（名詞）」は「兵庫＜地名：単独＞／県内＜ラベルなし＞」のように推定される。推定に用いる情報は、着目している単語を含み前後２単語ずつ合計５単語に関する見出し語、品詞の情報である。各ラベルの尤もらしさを確率として計算し、１個の文全体における確率の積の値が高くなり、かつラベルとラベルの間の連接規則を満たすように付与するラベルを決める。１個の文における最適解の探索には各処理段階における最適解をすべて保持する公知のアルゴリズムを用いていることができる。

（３）システムがよく生じる誤りについてその誤りを訂

正する書き換え規則を予め規則性の１つとして用意しておき、これを後処理に用いる。例えば、「兵庫＜地名：単独＞／県内＜ラベルなし＞」は「兵庫県＜地名：単独＞／内＜ラベルなし＞」のように書き換えられる。

（４）最後にこの結果から「兵庫県」を地名として抽出する。

本発明における手法によると、人間のパフォーマンスの９割程度の精度で固有表現を抽出でき、従来に比して効果的な固有表現の抽出が可能となった。

【００３５】以上のように本発明では、解析から生成に亙るテキスト処理を、最大エントロピーモデルを用いた学習という一貫した枠組みで処理をしている。そして、解析過程、すなわち形態素解析（単語の切り出し、品詞推定）、係り受け解析や、固有表現抽出を行う意味解析システムから、生成（語順の学習と決定）に至るまでの各処理を、予め解析済みのテキストを用いた学習によって実現する。さらにそれらを繰り返して実行することによって、少ない学習データにもかかわらず、大量の人手をかけて作成される規則に基づく方法に近い精度を実現でき、コストの抑制だけでなく、幅広い文章に対応可能なテキスト処理方法を提供することができる。これら技術は、自動翻訳技術や、テキストの要約技術に用いるだけでなく、例えば、コンピュータにおけるかな漢字変換等、いかなる言語処理にも適用することが可能である。

【００３６】

【発明の効果】本発明は、以上の構成を備えるので、次の効果を奏する。請求項１に記載のテキスト処理方法によると、解析過程及び生成過程を互いに繰り返して実行することによって、学習を行う解析済みテキストが少ない場合であっても、効果的に最大エントロピーモデルを用いた学習を行うことができ、高精度なテキスト処理方法を提供することができる。これによって、コストの低廉化と共に、高機能化を図ることができる。

【００３７】請求項２に記載のテキスト処理方法によると、形態素解析に最大エントロピーモデルを用いることができるので、請求項１に記載の循環的な学習に好適であり、コンピュータにおける処理に馴染みやすい。これによって、本発明におけるテキスト処理方法はより高精度化を図ることができ、処理の高速化にも寄与する。

【００３８】請求項３に記載のテキスト処理方法によると、係り受け確率を定数的に求めることができるので、＊より高精度な係り受け関係を導出することができ、ひいては高精度なテキスト処理方法に奉仕する。

【００３９】請求項４に記載のテキスト処理方法によると、１個の文全体について全ての係り受け関係の確率を求めるので、文全体として最適な係り受け関係を導出することができ、高精度な係り受け解析が可能となる。これにより高精度なテキスト処理方法に寄与する。

【００４０】請求項５に記載のテキスト処理方法によると、学習によって語順モデルを構築するので、学習を行う解析済みテキストが少ない場合であっても、効果的に学習を行うことができ、高精度なテキスト処理方法を提供することができる。

【００４１】請求項６に記載のテキスト処理方法によると、請求項５の方法により構築された語順モデルを用いることができるので、最適な語順の決定を効果的に行うことができる。

【００４２】請求項７に記載のテキスト処理方法によると、固有表現の抽出処理を行うので、形態素解析の精度向上に寄与し、ひいては高精度なテキスト処理方法が実現できる。

【図面の簡単な説明】

【図１】本発明によるテキスト処理方法の説明図

【図２】形態素解析の説明図

【図３】係り受け確率の算出実施例における各確率一覧図

【図４】係り受け確率の算出実施例

【図５】語順の学習生成における順序が適切である確率の計算例

【符号の説明】

１　テキスト処理方法
１０　テキスト
１１　統語構造
１２　形態素解析処理
１３　係り受け解析処理
１４　語順の学習生成処理
１５　固有表現抽出処理
３１　後文節を越えて先にある文節に係る確率
３２　後文節に係る確率
３３　後文節との間にある文節に係る確率
５１ないし５６　係り文節の語順の並べ替え例

---